Harmati Attila

Adatbányászat üzleti szemmel
II. rész

A cikk az adatbányászat technológiájának gyakorlati alkalmazását szemlélteti egy – a szerző által készített – adatbányászati projekt bemutatása révén. Az esettanulmány elkészítésének célja egy osztályozási modell kialakítása volt logisztikus regressziós modellek, döntési fa és neurális háló felhasználásával. A megfelelő eljárás megtalálásával az adatállományt nyújtó vállalat egy jövőbeli direktmarketing-akcióhoz kapcsolódóan tudatosan jelölheti ki azon ügyfelek körét, akik egy személyes levélre nagy valószínűséggel kedvezően reagálnának. A potenciálisan kedvező ügyfelek ezáltal hatékonyabban érhetőek el a véletlen kiválasztáshoz képest, így növelve a vállalat hatékonyságát és eredményességét. Az eredmények általánosítása alátámasztja az adatbányászat üzleti célú felhasználásának számos előnyét.[1]

JEL kód: C25, C44, C45, C49, C88

Kulcsszavak: *adatbányászat, osztályozás, direktmarketing, SAS®, SEMMA*

Bevezetés

A tanulmány jelen részében bemutatásra kerülő esettanulmány vezérfonalát a projekt elkészítéséhez felhasznált SAS® Enterprise Miner™ adatbányászati szoftver elemzési logikája alkotja. Ez öt fő lépésből áll, mely a mintavételezést, a feltárást, a módosítást, a modellépítést és az értékelést jelenti. Ezen lépések angol megfelelőiből, azaz a *Sample*, az *Explore*, a *Modify*, a *Model* és az *Assess* szavakból képzett SEMMA akronímával egyetlen szóba foglalható össze a projektkészítés folyamata. A szoftver segítségével elkészíthető adatbányászati projekt során a felhasználható eszközök igen gazdag választéka az egyes lépéseknek megfelelő csoportosításban találhatóak meg, melyek ezenfelül az egyéb eszközöket is felvonultató *Utility* kategóriával egészülnek ki *(SAS 2006)*.

Sample mint mintavételezés

Egy adatbányászati projekt első lépése az elemezni kívánt adatok részének vagy egészének munkafolyamatba importálása, illetve annak több részre történő felosztása a további lépések hatékonyabbá tétele érdekében. Ezeket az Enterprise Miner™-ben a mintavételezésről

Harmati Attila a Debreceni Egyetem Közgazdaságtudományi Karának végzett hallgatója. E-mail: harmatiati@gmail.com

[1] *A szerző köszönetet mond az elemzéshez felhasznált adatállományt biztosító vállalatnak és segítőkész munkatársainak, valamint dr. Ispány Mártonnak és Varga Sárának a tanulmány kapcsán nyújtott hasznos tanácsaikért.*

elnevezett eszközcsoport segítségével lehet megtenni, mely lépésben mindenekelőtt a használandó adatállományt kell a projektbe helyezni *(SAS 2006)*.

Az esettanulmány elkészítéséhez egy magyarországi telekommunikációs vállalat által nyújtott, 7500 egyedből álló minta[2] került felhasználásra. A minta eredeti formájában egy egyedazonosító mellett 1 területi, 14 időbeli és 109 tárgyi, utóbbin belül 23 minőségi és 86 mennyiségi változót tartalmazott, melyek az ügyfelek demográfiai és szociológiai jellemzői mellett azok tranzakciós magatartását reprezentálják egy három hónapos adatsor átlagolása révén. Ezen ügyféljellemzők mellett két változó egy korábbi, telefonegyenleg-feltöltésre ösztönző direktmarketing-kampány pozitív, illetve negatív hatását, és az esetleges pozitív hatás összegszerű mértékét is bemutatja az egyes ügyfelekre vonatkozóan.

Az adatok könnyebb kezelhetősége érdekében – azok importálása előtt – a szükséges változókra a Microsoft® Office Excel® programban kódolást alkalmaztunk, mely révén a területi és minőségi változók zömét numerikussá alakítottuk. Ezzel párhuzamosan az időbeli változókat különbségképzés révén eltelt napokká transzformáltuk át. Ezek eredményeként és 4 feleslegessé vált időbeli változó[3] mellőzésével a minta elemzésre felkészített állapotában az egyedazonosító mellett 118 mennyiségi változót és csupán 2 minőségi változót tartalmaz.

Az adatok Enterprise Miner™ projektbe illesztését a metaadatok létrehozása kell hogy kövesse a *Metadata* menüpont által *(SAS 2006)*. A metaadat *Berry – Linoff (1997)* megfogalmazásával élve adat az adatról, mely az adattáblában szereplő adatok fizikai szerkezetét, tehát az adatbázis vázlatát jelenti. Egy szemléletesebb definíció szerint ez „egy másik adatot leíró adat, amely összefoglalja az adat használatára vonatkozó összes fontos tényt" *(Márkus 1994:23.)*. Ilyen tényeket jelent az egyes változókra vonatkozóan azok mérési szintje, illetve azok elemzésben betöltött szerepe. A program ezeket megvizsgálja, azonban szükség esetén korrekciós lehetőség is van. Módosításra volt szükség például a korábbi marketingkampány kimenetének milyenségét leíró változó célváltozóvá tétele[4], illetve az egyenlegfeltöltés összegét leíró változó mellőzésének beállítása érdekében.

Az adatokban rejlő információk feltárási folyamatának megkezdése előtt az adatállományt a *Data Partition* elnevezésű menüpontban három részre osztottuk fel: egy tanuló-, egy érvényesítő- és egy tesztelőállományra. Általánosságban elmondható, hogy a modellek előzetes felépítése a tanulóállomány segítségével történik. Ezt támogatandóan, pontosabban felügyelendően a modelleket közvetve lehet finomítani és időben leállítani az érvényesítőállomány használatával, míg a tesztelőállománnyal a modellek értékelése történik *(SAS 2006)*. Ezek arányait 40, 30, 30%-ban határoztuk meg, melyekhez a szükséges megfigyelési egységeket – azaz a rekordokat – a program rétegzett mintavételi móddal választotta ki, így biztosítva a három állományban a célváltozó kimenetének azonos arányát *(SAS 2008)*.

[2] Hangsúlyozni szükséges a minta teljes mértékű anonimitását.

[3] Feleslegessé vált az adatfelvétel egységes időpontját bemutató változó, valamint a szerződés aktiválásának, a hűségszerződés létrehozatalának és lejártának dátumát bemutató változó, mivel az eredeti minta tartalmazta a szerződés élettartamát, a hűségszerződés hosszát, illetve az abból még hátralévő időt reprezentáló változókat.

[4] A mintavételezés módjáról a célváltozó ismeretében lehet érdemben szólni. Ez a koncentrált kiválasztás módszere volt a pozitívan reagáló ügyfelek megfelelően magas arányának biztosítása érdekében.

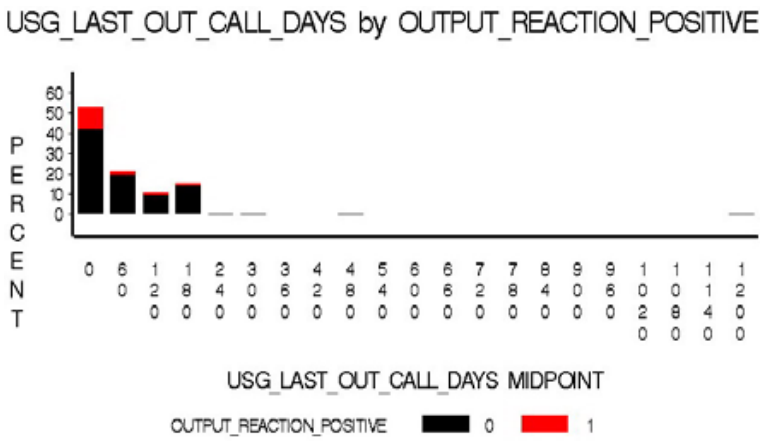
Explore mint feltárás

A hipotézisek felállítása, valamint a modellek helytálló felépítése érdekében az elemzőnek meg kell ismernie az elemzés tárgyát képező adatokat, mely folyamat során az alapvető összefüggések és trendek feltárására van szükség *(SAS 2006)*.

A feltárás kategóriájában található *MultiPlot* elnevezésű eszköz segítségével grafikus megjelenítésben vizsgálhatók az egyes változók osztályonkénti gyakoriságai, például oszlopdiagramok formájában *(SAS 2006)*. Erre látható példa az 1. ábrán, mely az utolsó indított hívás óta eltelt időt reprezentáló változó tanulóállományra eső részének gyakorisági eloszlását szemlélteti.[5]

1. ábra

Az utolsó indított hívás óta eltelt idő tanulóállomány-beli gyakorisági sora



Megjegyzés: Az abszcisszán az eltelt idő napokban, az ordinátatengelyen az adott időszakhoz tartozó gyakorisági érték százalékában van megadva. A fekete szín a célváltozó adott időtartamhoz tartozó negatív kimenetét, a piros szín pedig annak pozitív kimenetét érzékelteti.
Forrás: saját készítés.

Egy másik feltárást szolgáló, *StatExplore* elnevezésű többfunkciós eszközzel a változók eloszlása és alapvető statisztikai mutatói ismerhetők meg *(SAS 2006)*. Ezek hozzásegítik az elemzőt a további kutatás céljából kijelölendő változók megtalálásához, előzőeken felül például a korrelációs együtthatók meghatározása által, mellyel a redundáns változók felismerhetővé válnak.

Alapstatisztikai vizsgálatok

A projekt ezen szakaszában az elemzés látókörének szélesítése érdekében a minta alapvető tulajdonságainak részletes megismeréséhez a SAS® Enterprise Guide™ statisztikai szoftver által nyújtott lehetőségeket is felhasználtuk.

[5] *Az osztályhatárok önkényesen kerültek meghatározásra.*

A mennyiségi változókra vonatkozóan az alapvető helyzet-, szórás- és alakmutatók kiszámítása alapján az ügyfelekről általánosságban elmondható, hogy 98,7%-uk magánelőfizető, akik mindegyike előre fizet a telekommunikációs szolgáltatásért, így – az esettanulmány terminológiájához igazodva – a prepaid előfizetési kategóriába tartozik[6]. Az ügyfelek többsége olyan előfizetési díjcsomaggal rendelkezik, mely által a nap bármely szakában bármely vezetékes, illetve a szolgáltató számára konkurens hálózatba egységes díjért tud telefonálni. Ebből a tranzakciós szokásokat reprezentáló változók megismerése előtt – az ügyfelek racionális magatartását feltételezve – az a következtetés vonható le, hogy azok kommunikációs partnereinek többsége valószínűleg nem az érintett szolgáltatóhoz tartozik.

A demográfiai jellemzők alapján megállapítható, hogy a mintában szereplő ügyfelek 56,4%-a férfi, átlagéletkoruk 39,5 év, jellemzően egyedülállóak és középfokú végzettséggel rendelkeznek.

A szolgáltatóval való szerződéskötés helyéből arra lehet következtetni, hogy a mintában szereplő ügyfelek túlnyomó többsége vidéki lakos. Ezek a szerződések jellemzően hűségszerződés nélküliek – az ügyfelek mindössze 0,25%-nak van hűségszerződése –, és átlagosan 32 hónapos, így frissnek egyáltalán nem tekinthető kapcsolatot jeleznek a szolgáltató és ügyfele között.

Az igénybe vett szolgáltatásokat szemlélve az látható, hogy a speciálisnak nevezhető, kiegészítő szolgáltatások – úgymint a WAP, a roaming és az e-mail – az ügyfeleknél jellemzően nincs aktiválva, bár a szintén ebbe a kategóriába tartozó MMS-szolgáltatás az ügyfelek 18,9%-ánál aktív.

Az egyenlegfeltöltések átlagos száma a három hónapos időtartamra vonatkozóan 0,046, ennek megfelelően ezer ügyfélből egy-egy feltöltést átlagosan 46 fő hajtott végre ez időszak alatt. Az utolsó feltöltési alkalom óta az adatfelvétel időpontjáig ügyfelenként átlagosan 173 nap telt el, a telefonálási aktivitást tekintve pedig a három hónapos használat összege ügyfelenként átlagosan 958 forint. Míg a kimenő hívások hossza átlagosan 16,7 perc, és az utolsó indított hívás óta átlagosan 56 nap telt el, addig a bejövő hívások hossza ügyfelenként 55,4 perc, és az utolsó fogadott hívás óta 120 nap telt el. Az üzenetküldési aktivitást szemlélve az látható, hogy három hónap alatt átlagosan 2,4 darab SMS-üzenetet küldtek az ügyfelek, és az utolsó üzenet küldése óta 315,9 nap telt el. A fogadott üzenetek száma ügyfelenként átlagosan 20,2 darab, és az utolsó fogadása óta átlagosan 63,5 nap telt el.

A direktmarketing-kampány hatását bemutató változók értékeiket annak megfelelően veszik fel, hogy a kampány egy adott személynél elérte-e a célzott hatást, és ha igen, akkor milyen összeggel történt az egyenlegfeltöltés. A hatás milyenségét bemutató bináris változó számtani átlaga 0,1463, így megállapítható, hogy a megkeresettek 14,63%-ára volt hatással a kampány, 85,37%-uk pedig nem reagált. A hatás mértékét bemutató változó számtani átlaga 809 forint, szórása pedig 2806,14 forint. Ez azt jelenti, hogy a 809 forintos átlagos egyenlegfeltöltéstől a megkeresettek átlagosan 2806,14 forinttal eltérő összeget költöttek mobilszolgáltatójuknál. De mivel a negatív egyenlegfeltöltés a gyakorlat szempontjából használhatatlan információ, ezért a figyelem a változó relatív szórására kell hogy koncentrálódjon, ami 3,468, azaz 346,8%. Ebből egyértelműen látható, hogy a pozitívan

[6] *Ebben a perspektívában az ügyfelek két kategóriája különböztethető meg, a prepaid és a postpaid kategória. Előbbi esetén adott ügyfél az igényelt szolgáltatásokat előre, egy összegben fizeti ki, utóbbi esetben pedig a szolgáltatások árát nagyobb időközönként, például havonkénti gyakoriságban törleszti (Ary – Imre 2006).*

reagálók esetében jelentősen eltérő összegekkel valósult meg a kívánt hatás, és erről tanúskodik az igen extrém, 71 800 forintos maximum is.

Függőség-, függetlenség-vizsgálat

Bármely két változó közti kapcsolat vizsgálata alapjában az azok közötti összefüggések, okokozati kapcsolatok és ható tényezők feltárását és elemzését jelenti *(Hunyadi – Vita 2004)*. Előbbiek részletes megismerése érdekében ismét az eszközök széles eszköztárát felvonultató Enterprise Guide™ szoftvert használtuk fel.

A változók közti kapcsolatok számszerűsítésére mérési szintjüknek megfelelően másmás módszerek és mutatók alkalmazandók, a mérési skálák különböző kombinációi esetén pedig alapkövetelmény, hogy a gyengébb skálának megfelelő kapcsolati típushoz tartozó vizsgálati módszer kerüljön felhasználásra *(Hunyadi – Vita 2004)*.

A korábbi direktmarketing-kampány hatását bemutató változót leginkább befolyásoló tényezők megtalálása érdekében a kampány hatására történt egyenlegfeltöltési összegeket bemutató – arányskálán mért – változó[7] és a 21 nominális, a 93 arány-, illetve a 2 sorrendi skálán mért változó között fennálló kapcsolatok mértékét vizsgáltuk[8]. Azonban a kapcsolatok erősségét reprezentáló mutatók megismerése önmagában egyetlen elemzőt sem győzhet meg egyértelműen az egyes változók mellőzhetőségéről, mivel azokat igen jelentősen befolyásolhatják a kiugró, extrém értékek. A döntéshez egy magasabb szintű statisztikai eljárásra, a hipotézisvizsgálatra is szükség van.

Az egyes változók és a direktmarketing-kampány hatását bemutató változó közötti kapcsolatok szignifikánsságának megerősítésére alkalmazott hipotézisvizsgálat során a minimálisan elvárható 95%-os biztonsági szintet, azaz 5%-os szignifikanciaszintet tartottuk szem előtt.

A nominális változók közt fennálló kapcsolatok szignifikánssága a Pearson-féle khinégyzet-próba által vizsgálható *(Hunyadi – Vita 2004)*. Ettől eltérő módszer alkalmazandó az arány-, illetve sorrendi skálán mért változók kapcsolatának tesztelése, a vizsgálat ez esetben a Fisher-féle *z*-transzformáció felhasználásával végezhető el (lásd *Fisher 1915*).

A próbák eredményéül adódó p-értékek[9] alapján mindössze 7 nominális és 66 arányskálán mért változóról állítható 95%-os biztonsággal, hogy azok szignifikáns kapcsolatban állnak a direktmarketing-kampány eredményét reprezentáló változóval. Az eredmények tükrében összegzésként megállapítható, hogy a vizsgált 116 változóból 73 bizonyult szignifikánsnak a kampányreakció alakulásának tekintetében, mely a figyelem koncentrálása révén a további elemzés szempontjából igen kedvezőnek tekinthető.

[7] Az egyenlegfeltöltési összegeket reprezentáló változó 0 értéket ugyanazon ügyfelek esetén vesz fel, mint a kampányreakciót bemutató bináris változó.

[8] Nominális skálán mért változók esetén a Cramer-féle asszociációs együtthatót, arányskálán mért változók esetén a Pearson-féle lineáris korrelációs együtthatót, sorrendi skálán mért változók esetén pedig a Spearman-féle rangkorrelációs együtthatót alkalmaztuk.

[9] Fontos kiemelni, hogy a Fisher-féle z-transzformációt alkalmazó próbák esetén a kétoldali alternatív hipotézis miatt a próbafüggvény értékéből számított szignifikanciaszint kétszerese a p-érték.

Modify mint módosítás

Az adatok megismerése után következő lépés a változók szükség szerinti módosítása, például az extrém értékek[10] kezelése, valamint a változók átalakítása a jobb, tehát pontosabb és nagyobb teljesítményű modellek megalkotásának elősegítése érdekében *(SAS 2006)*. A változók átalakítása két eltérő módon valósítható meg. Az egyik út a feltárás folyamatában relevánsnak ítélt, így elemezni kívánt változók kiugró és hiányzó értékeinek kezelésén túl új változók létrehozását jelenti a meglévők transzformálása által. A másik módszer a feltárás eredményéül adódó változószelekció figyelmen kívül hagyásával új változók létrehozását, majd azok szelektálását jelenti az egyes változókhoz tartozó értékek csoportokba foglalása által *(SAS 2008)*.

Változók módosítása

Az adatok módosításának első módszere során mindenekelőtt a szélsőséges értékeket felvonultató változók kezelése szükséges, melyre a *Filter* elnevezésű eszköz alkalmazható. Ezen értékek tanulóállomány-beli jelenlétének feltárására különböző lehetőségek állnak rendelkezésre, például a nagy szóródással rendelkező vagy egyszerűen a ritka értékek megkeresése *(SAS 2006)*. Előbbi módszert az intervallumváltozókra alkalmaztuk oly módon, hogy az átlagtól háromszoros szórásnyi távolságon túl eső értékeket tekintettük extrémnek. A kategóriaváltozókra a ritka értékek kritériumát használtuk fel azon feltételezés mellett, hogy az 1%-tól kisebb relatív gyakoriságú értékeket ítéltük szélsőségesnek. Az így meghatározott kiugró értékeket a további vizsgálatok torzításának csökkentése érdekében kiszűrtük.

Az extrém értékek eltávolítása után a hiányzó értékek pótlása szükséges, mivel egyes modellépítési eljárások – például a regressziók – nem alkalmasak azok kezelésére. Ezt az *Impute* elnevezésű eszköz segítségével lehet megtenni, mely több választási lehetőséget is felvonultat *(SAS 2006)*. Az intervallumváltozók hiányzó értékeinek pótlására azok egyszerű számtani átlagát, a kategóriaváltozók szükséges pótlására pedig azok leggyakrabban előforduló értékeit használtuk fel. A hiányzó értékek pótlásán kívül ebben az eszközben adható meg a hiányzó értékek azon maximális aránya, mely felett a szoftver egy adott változót a továbbiakban mellőzendőnek ítél *(SAS 2008)*, ezt 50%-ban határoztuk meg.

A kiugró értékek kiszűrése és a hiányzók pótlása után az új változók létrehozása következhet a már meglévők transzformálása által a *Transform Variables* menüpontban[11]. A transzformáció használata által stabilizálható az egyes változók varianciája, javítható normalitása, és eltávolítható nem-linearitása *(SAS 2006)*. Intervallumváltozók esetén a logaritmikus transzformációt preferáltuk annak kedvező tulajdonságai miatt, kategória-változók esetén viszont nem alkalmaztunk transzformációt.

[10] A szélsőséges értékek alapvetően a valóságnak megfelelő, de szokatlanul extrém, valamint a hibás adatokat jelentik. Utóbbi származhat például gépelési pontatlanságból.

[11] A menüpont lehetőséget nyújt változóbővítésre is, például arányváltozók képzése által.

Az adatok módosítására használható második módszer az *Interactive Binning* menüpont segítségével hajtható végre. Ez esetben a modellezésre felhasználandó változók kiválasztása a Gini-statisztikára[12] alapozva történik meg. A mutató kiszámítása előtt a változók értékeit csoportokba kell sorolni oly módon, hogy adott változóhoz képzett csoportok eseményrátái[13] minél eltérőbbek legyenek *(SAS 2008)*. Ennek érdekében csoporthatárokként a változók kvartiliseit határoztuk meg. Az így létrehozható változószelekció révén az osztályozás, így az előrejelzés erősebbé válhat, a modellek túlillesztése pedig elkerülhető. A kategorizálást – mint transzformációt – kategóriaváltozóknál akkor hasznos alkalmazni, ha az sok értéket vehet fel, és számos értékhez tartozó gyakoriság elenyésző szintű. Intervallumváltozók esetén ez a transzformáció akkor előnyös, ha annak kapcsolata a célváltozóval nem lineáris, és egyéb, például logaritmikus transzformációval sem tehető azzá *(SAS 2008)*.

A felhasznált Gini-statisztika az inputváltozók célváltozóra vonatkozó szeparálási képességét, tehát a diverzitást méri oly módon, hogy értékének egyre nagyobb mértéke egyre nagyobb diverzitást, tehát a változóhoz képzett osztályok egyre eltérőbb eseményrátáit mutatja *(Lucas 2004)*. Ennek figyelembevételével a változók közül azokat tekintettük relevánsnak, melyek Gini-statisztikája meghaladta a 20-at. Ezáltal a 118 inputváltozóból már csak 29 maradt informatív a további elemzés szempontjából.

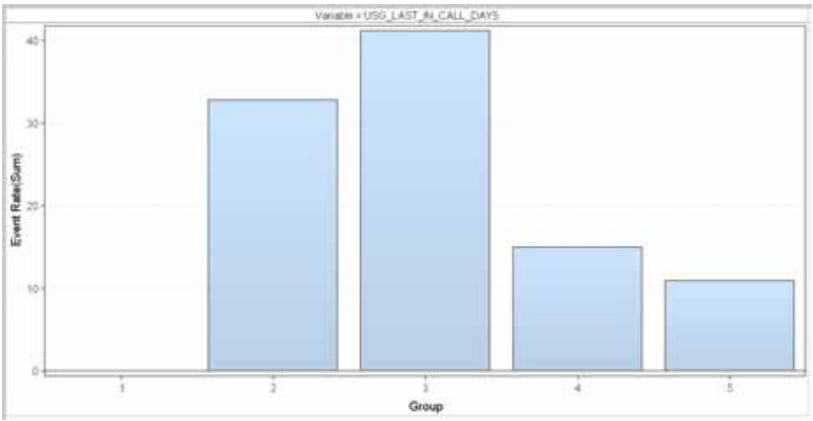
Az eljárást szemléltető példa kiindulópontja az 1. ábrán bemutatott gyakorisági eloszlás, melyen jól látható, hogy 180 nap eltelte után már nem számottevő az egyes kategóriákba eső elemek száma, így ez esetben a csoportosítás hasznosnak tűnik. A változó kvartilisei a 6., 27. és 98. nap, melyek helyett azonban a differenciáltabb pozitív célváltozó-kimenet érdekében kisebb módosítást követően csoporthatárként az 5., 24. és 92. napot definiáltuk. Az első csoportban a hiányzó értékek szerepelnek, a másodikban az 5 naptól kisebb értékű megfigyelések, a harmadikban az 5, illetve az 5. és 24. nap közti, a negyedikben a 24, illetve a 24. és 92. nap közti, az ötödikben pedig a 92 naptól nagyobb értéket felvevő megfigyelések találhatók. Az elkészített csoportosítás grafikus megjelenítése a 2. ábrán látható.

[12] $Gini = 1 - \left(\frac{2\sum_{i=2}^{m}(n_i^{pozitív} * \sum_{1}^{i-1} n_i^{negatív}) + \sum_{i=1}^{m}(n_i^{pozitív} * n_i^{negatív})}{N^{pozitív} * N^{negatív}}\right) * 100$, *ahol i=1,…,m az adott változóban képzett csoportok számát jelöli,* $n_i^{pozitív}$, *illetve* $n_i^{negatív}$ *az i-dik osztályba tartozó azon egyedek száma, melyekre a kívánt esemény teljesül, illetve nem teljesül,* $N^{pozitív}$, *illetve* $N^{negatív}$ *az adott változóhoz tartozó összes egyed száma, melyek esetén teljesül, illetve nem teljesül esemény (Lucas [2004]).*

[13] *Az eseményráta egy esemény relatív gyakorisága, így a következő képlettel definiálható:* $\frac{n^{pozitív}}{n^{összes}}$, *ahol* $n^{pozitív}$ *azon megfigyelések számát jelöli, melyekre a kívánt esemény teljesül,* $n^{összes}$ *pedig az összes megfigyelés száma.*

2. ábra

Egy változó értékeinek kategorizálása



Megjegyzés: Az abszcisszán a változó értékeihez képzett csoportok, az ordinátatengelyen az adott csoporthoz tartozó, a célváltozó pozitív kimeneti arányát reprezentáló mutató látható.
Forrás: saját készítés.

Üzleti szempontok integrálása

Egy osztályozási feladat során két fajta hiba követhető el. Ezek az esettanulmányra aktualizálva úgy fogalmazhatók meg, hogy elsőfajú hiba esetén a megkeresésre nem reagálók csoportjába kerülnek besorolásra bizonyos ügyfelek, akik valójában pozitívan reagálnának, másodfajú hiba esetén pedig nem reagáló ügyfelek kerülnek a pozitívan reagálók csoportjába. Előbbi egy potenciális ügyfél elvesztését, utóbbi felesleges megkeresési költséget jelent. A valós költség-haszon értékeket tükröző modellek felépítése érdekében az egyes döntések következményeit definiálni kell *(SAS 2006)*, amit az 1. táblázat foglal magában.

1. táblázat

Döntési mátrix

	Levél elküldése	Nincs levélküldés
Pozitív reakció	13,62 euró	0 euró
Nincs reakció	–0,77 euró	0 euró

Forrás: saját készítés.

A táblázat a feltüntetett szituációkhoz tartozó profit-következményeket mutatja be azon feltételezés mellett, hogy egy levél előállítása és postázása nagyságrendileg 200 forintba kerül, valamint az adatbázisban szereplő, pozitívan reagáló ügyfelek egyenlegfeltöltéseinek mediánja 3750 forint. Ez egy megkeresés után 3550 forint profitot jelent. Ezeket, illetve a továbbiakban használandó összegeket az esettanulmány aktualitásának konzerválása érdekében a Magyar Nemzeti Bank 2008. november 26-án érvényes hivatalos devizaárfolyama alapján 260,68 forint/euró ráta alapján váltottuk át euróra *(MNB 2008)*.

A modellalkotás realitásának megőrzése érdekében a pozitív reakciók a priori valószínűségét 4%-ban, a reakció elmaradásának valószínűségét 96%-ban határoztuk meg.[14]

Model mint modellépítés

Az adatbázis előkészítése után következhet az adatbányászati projekt egyik leglátványosabb eleme, a modellépítés. Ennek során a feladat az adatok elemzése analitikus eszközökkel, például regressziók, döntési fák, neurális hálók és ezek kombinálása által létrehozott modellek készítése révén, melyek jó becsléseket, illetve előrejelzéseket képesek adni a célváltozó kimenetével kapcsolatban *(SAS 2006)*.

Az elkészített modellek teljesítményét két-két ábrával szemléltetjük, melyek a válaszarány és a válaszadók koncentrációs együtthatójának kumulatív mérőszámaira összpontosítják a figyelmet. A válaszarány adott kiválasztási arány mellett az ügyfelek azon hányadát mutatja meg, akik a vizsgált modell alapján – a populáció összetételét változatlannak feltételezve – egy megkeresésre várhatóan pozitívan reagálnának. Ehhez a fogalomhoz kapcsolódóan feltétlenül meg kell említeni a lift-érték mutatót, mely azt mutatja meg, hogy a célközönség adott százalékának a modell alapján történő megkeresésével elérhető válaszarány hányszorosa egy véletlen kiválasztásból eredő válaszaránynak, amit 4%-ban definiáltunk. A modellek teljesítményének teljesebb körű megismerését teszi lehetővé a válaszadók koncentrációs együtthatója, mely az összes potenciálisan jó ügyfél adott kiválasztási arány mellett található százalékos arányát mutatja meg *(Coppock 2002)*.[15] A modellek teljesítményét külön-külön bemutató ábrákon a kék színnel jelölt görbék a tanulóállományra, a piros színnel jelölt görbék pedig az érvényesítőállományra vonatkozó adatokat szemléltetik, így a figyelem a piros görbékre kell hogy koncentrálódjon.

Logisztikus regressziós modellek

A regressziós modellek felállítására alkalmas *Regression* elnevezésű eszközt az adatokra illesztendő lineáris és logisztikus regressziós modellek készítésére lehet alkalmazni *(SAS 2006)*. A tanulmány első részében már kitértünk rá, hogy logisztikus regressziós modellek segítségével egy adott esemény bekövetkezési valószínűsége becsülhető. Ennek megfelelően az ügyfelek direktmarketing-kampányra adott reakciói bekövetkezésének valószínűségét függő változóként használó logisztikus regressziós modelleket készítettünk.

A függő változó intervallumbeli korlátozottságát feloldandóan a logit transzformációt alkalmaztuk, a paraméterek becslésére pedig a megfigyelt esemény valószínűségét maximalizáló maximum likelihood módszert. A modellek felépítésének módjaként a Wald-teszt és a likelihood-hányados teszt alkalmazása során szignifikánsnak tűnő változókat a modellbe egyenként bevonó, illetve azokat szignifikánsságuk megkérdőjelezésekor a modellből eltávolító stepwise eljárást használtuk, modellkritériumként pedig azt a módszert, mely a modell elkészítése során a várható profit maximalizálását, illetve az esetleges veszteség minimalizálását tartja szem előtt. A modellek használhatóságát reprezentáló

[14] Az a priori valószínűségek az adatállományt nyújtó vállalat szakemberei segítségével kerültek meghatározásra.

[15] A definiált mutatók az ábrák ordinátatengelyén – kumulatív értelemben – szerepelnek. Az abszcisszán az ügyfelek megcélozni kívánt százalékos aránya látható oly módon, hogy azok az aktuális modell előrejelzésére alapozva a megkeresésre történő pozitív reagálásuk becsült valószínűsége alapján csökkenő sorrendbe vannak állítva.

mutatók által a magyarázó változók hatásának mértékét a Nagelkerke-féle együtthatóval vettük figyelembe, a modellillesztés jóságát a Hosmer–Lemeshow-teszttel értékeltük.

Ezen beállítások elvégzése után a vizsgálat sokszínűségének fenntartása érdekében két módon készítettük el a regressziós modelleket, mely módszerek a felhasznált adatok módosításában különböznek. Az első esetben a szűrés, helyettesítés, transzformálás hármasát alkalmaztuk, mivel a regressziós modellek nem képesek megfelelően kezelni a hiányzó és extrém értékeket, valamint az erősen nem lineáris változókat *(SAS 2006)*.[16] A második esetben hasonló okok miatt az *Interactive Binnig* eszköz segítségével elvégzett csoportokra bontás és szelekció után készítettünk regressziós modellt.[17]

Az első regressziós modell által felhasznált tényezők fontossági sorrendben az alábbiak:

- Az egyik konkurens hálózathoz történő átirányítások számosságának logaritmusa,
- Indított hálózaton belüli hívások számának logaritmusa,
- Hívószámkijelzés aktivitása,
- Másodpercben mért WAP-használat hosszának logaritmusa,
- 1000–2000 forint[18] egyenlegű napok számának logaritmusa,
- 0–500 forint[19] egyenlegű napok számának logaritmusa,
- Utolsó fogadott SMS óta eltelt idő logaritmusa,
- Fogadott hétvégi hívások hosszának logaritmusa,
- Indított vezetékes hívások hosszának logaritmusa.

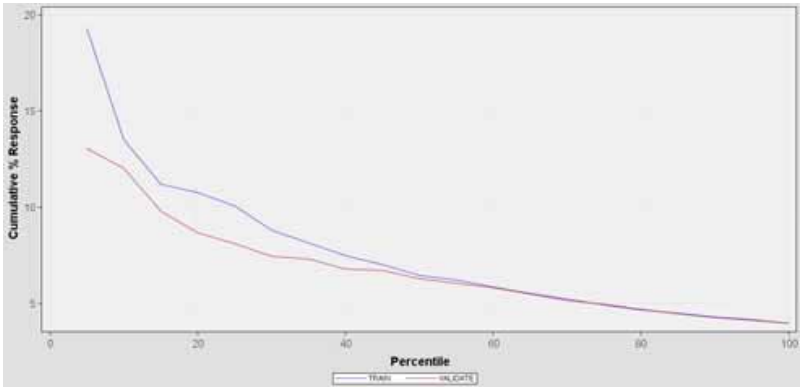
Az ezekből a változókból felépített logisztikus regressziós modell teljesítményét a korábban ismertetett, a kumulált válaszarányt és a válaszadók kumulált koncentrációs arányát reprezentáló görbék segítségével szemléltetjük.

A 3. ábráról a piros színnel jelölt érvényesítőállományra vonatkozóan leolvasható, hogy az első regressziós modell becslése alapján a legkedvezőbb várható reakciójúnak ítélt ügyfelek első 10%-ának megkeresése esetén 12,036%-os, 20%-uk megkeresésével pedig 8,693%-os pozitív válaszadás érhető el. Ezen adatok segítségével meghatározható a véletlenszerű megkeresés által elérhető válaszarányhoz viszonyított hatékonyságnövekedést bemutató kumulált lift-érték, mely jelen esetben 10% mellett 3,009-szeres, 20% mellett pedig 2,173-szeres javulást jelent.

[16] Az így elkészített modell az érintett ábrákon annak angol megfelelője alapján a Regression nevet viseli.
[17] Az így elkészített regressziós modellre az ábrákon a Regression (2) elnevezéssel utalunk.
[18] Az egyenlegösszeg 260,68 Ft/euró árfolyamon 3,83–7,67 eurónak felel meg.
[19] Az egyenlegösszeg 0–1,92 eurónak felel meg.

3. ábra

Az első regressziós modell kumulált válaszarány-görbéje

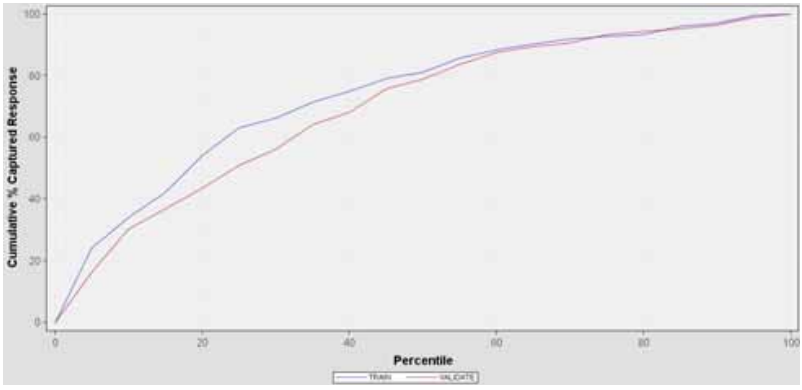


Forrás: saját készítés.

A 4. ábrán látható, hogy a legvalószínűbben várhatóan pozitív választ adó ügyfelek első 10%-ának megkeresése esetén a várhatóan pozitívan reagálók 30,091%-át, az első 20%-uk megkeresése mellett pedig 43,465%-át lehet elérni.

4. ábra

Az első regressziós modell koncentrációs görbéje



Forrás: saját készítés.

A második logisztikus regressziós modell által felhasznált tényezők[20] fontossági sorrendben az alábbiak:

- Utolsó indított hívás óta eltelt idő_2,
- Bejövő hívások száma_3,
- Utolsó indított hívás óta eltelt idő_3,
- Indított hálózaton belüli hívások száma_4,

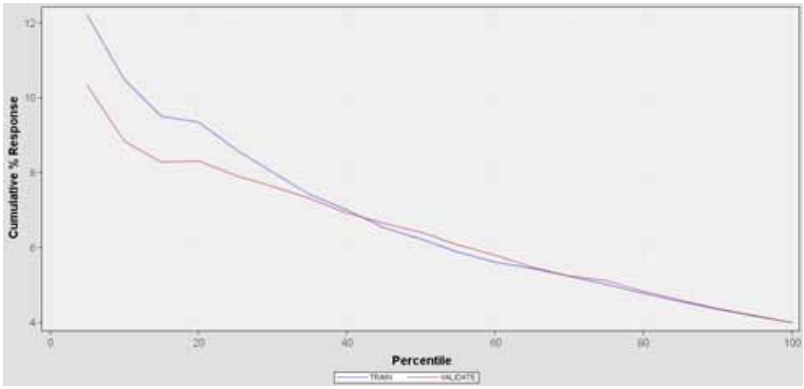
[20] A változók elnevezései mellett szereplő számok az adott változóhoz képzett, aktuális dummy változót jelölik.

- Bejövő hívások száma_4,
- Utolsó indított hívás óta eltelt idő_4,
- Indított hálózaton belüli hívások száma_3.

A modell teljesítménye az előzőekben bemutatott ábrákkal és értelmezésekkel ismerhető meg. Az 5. ábrán látható, hogy a második regressziós modell becslése alapján a várhatóan pozitívan válaszolók legjobb 10%-ának megkeresésével 8,828%-os, 20%-uk megkeresésével 8,305%-os pozitív válasz érhető el. Az ezekből az adatokból számított kumulált lift értékek tanúsága szerint ez a véletlenszerű megkereséshez képest 2,207-szeres, illetve 2,076-szeres javulást jelent.

5. ábra

A második regressziós modell kumulált válaszarány-görbéje

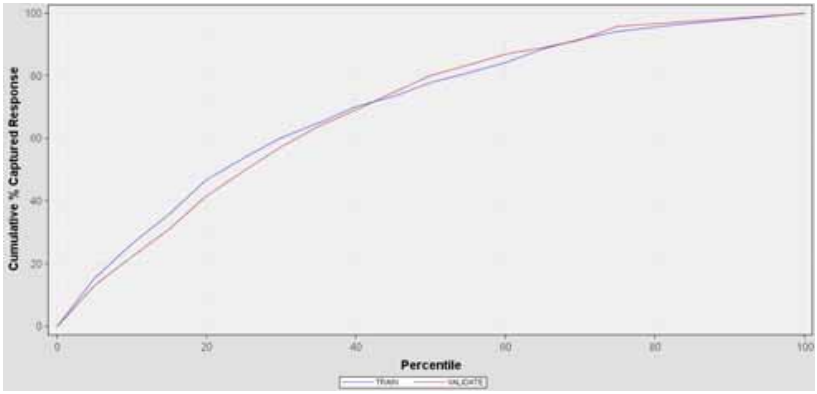


Forrás: saját készítés.

A 6. ábráról leolvasható, hogy az ügyfelek első 10%-ának megkeresése esetén a várhatóan pozitívan reagálók 22,072%-át, 20% esetén 41,526%-át lehet elérni.

6. ábra

A második regressziós modell koncentrációs görbéje



Forrás: saját készítés.

Döntési fa

A döntési fa *Decision Tree* elnevezésű eszköz segítségével történő elkészítéséhez az adatok módosítását mellőztük, mivel ezt a módszer nem igényli, a szélsőséges és hiányos adatokat is képes kezelni *(SAS 2006)*. A csomópontok szétvágásához felhasznált attribútumok kiválasztása a tanulmány első részében bemutatott, az információnyereséget az entrópia alapján megítélő módszerrel történt. A csomópontokhoz tartozó ágak maximális számát, azaz az egyes attribútumok lehetséges kimeneteit 2 darabban, a fa maximális mélységét, azaz a fa gyökerének teljes mértékű szétbontására használt attribútumok számát 6 darabban, a levelek, azaz a végső csoportok minimális elemszámát 65 darabban határoztuk meg. Az optimális fa megtalálásához azt a szabályt választottuk, mely végeredményként a legnagyobb átlagos profitot, illetve a legkisebb esetleges veszteséget ígéri.[21]

Az ezen beállításokkal elkészített döntési fa leegyszerűsített mását a 7. ábrán mutatjuk be. A fa piros színnel jelölt csomópontjai, illetve levelei a kedvezőtlen, a sárga színnel jelöltek a semleges, a zölddel jelöltek pedig a kedvező kimeneteket jelölik, a színátmenetek pedig értelemszerűen minőségi átmenetet reprezentálnak.

Az ábráról egyszerűen leolvashatók azok a „Ha-Akkor" szabályok, melyekkel egy újabb, osztályozási címkével nem rendelkező adatállomány kategorizálása is végrehajtható, ezáltal megkönnyítve a direktmarketing-levelek címzettjeinek kiválasztását. Az egyik legígéretesebb levélhez az alábbi módon lehet eljutni:

- Ha egy adott ügyfél esetén az utolsó fogadott hívás óta eltelt idő kisebb mint 21,5 nap,
- és az elmúlt három hónap alatt a használat összege nagyobb-egyenlő mint 1275 forint[22],
- és az utolsó feltöltés előtti egyenlege kisebb mint 2983 forint[23],
- akkor az ügyfél a megkeresésre nagy valószínűséggel pozitívan fog reagálni.

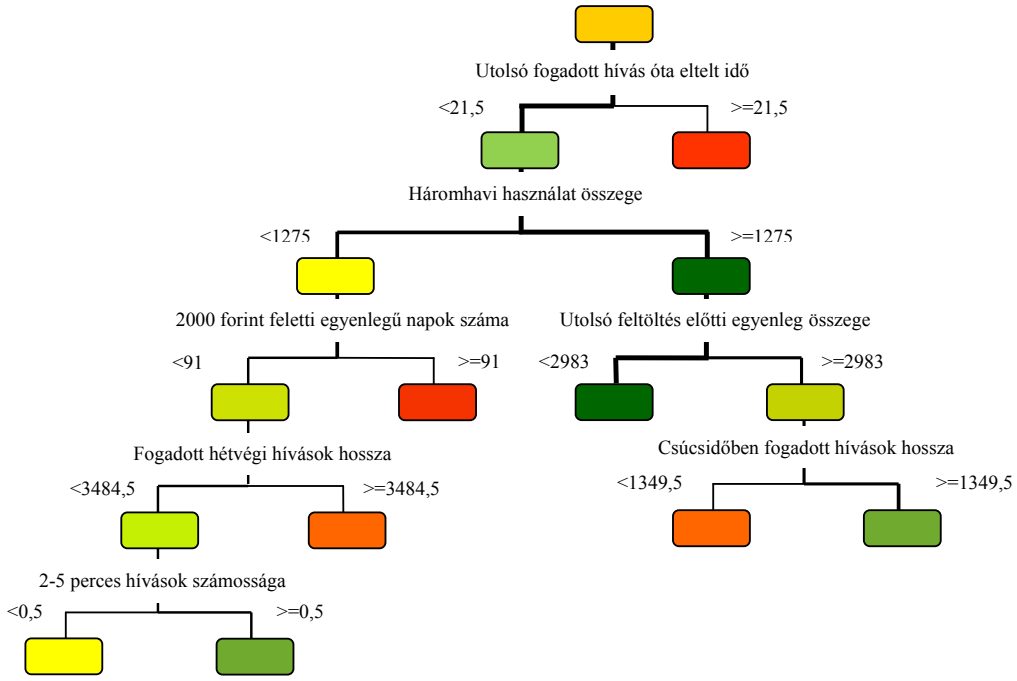
A terjedelmi korlátok miatt mellőzendőnek ítélt részletes ábra belső csomópontjai és levelei tartalmazzák az adott részhalmazban lévő elemek számát, azok kedvező, illetve kedvezőtlen kimeneteinek arányát és az egy főre jutó várható átlagos profitot mind a tanuló, mind az érvényesítőállományra. A fa imént bemutatott levelében szereplő adatokból a két állományra vonatkozó eredmények stabilitása mellett az is kiderül, hogy az ilyen tranzakciós jellemzőkkel rendelkező ügyfelek megkeresése esetén átlagosan 0,88 eurós profit realizálható, és a megkeresettek 35%-a reagál pozitívan.

A döntési fa teljesítményét bemutató 7. ábráról leolvasható, hogy annak becslése alapján az ügyfelek legjobb 10%-ának megkeresésével 11,471%-os, 20%-uk megkeresésével 9,92%-os pozitív válasz érhető el. Az ezekből számított kumulált lift-érték szerint ez a véletlenszerű megkereséshez képest 2,867-szeres, illetve 2,48-szoros hatékonyságjavulást jelent.

[21] A döntési fára az érintett ábrákon Default Tree-ként hivatkozunk.
[22] A használat összege 4,89 eurónak felel meg.
[23] A feltöltés előtti egyenleg 11,44 eurónak felel meg.

7. ábra

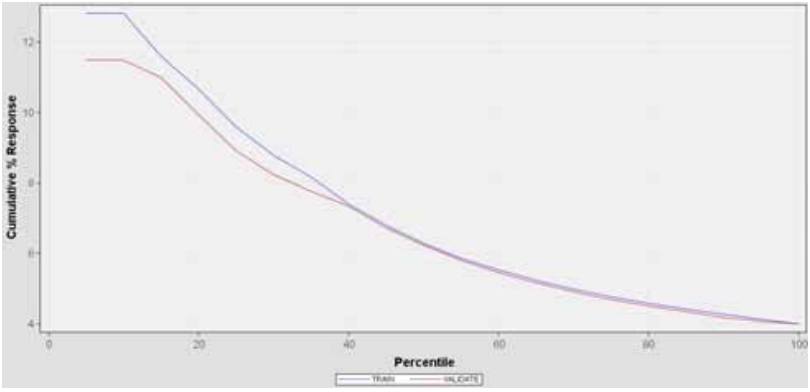
A döntési fa



Forrás: saját készítés

7. ábra

A döntési fa kumulált válaszarány-görbéje

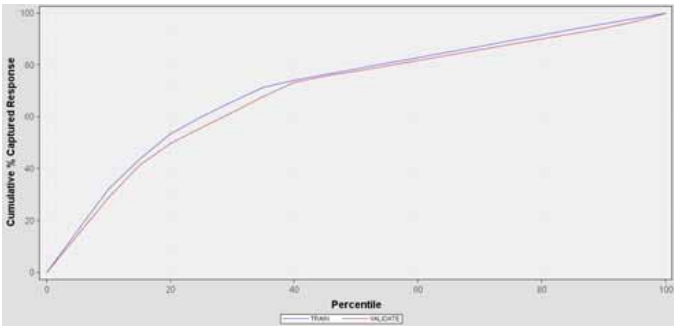


Forrás: saját készítés

A 8. ábra alapján megállapítható, hogy ez esetben az ügyfelek első 10%-ának megcélzása által a várhatóan pozitívan reagálók 28,678%-át, első 20%-uk megcélzásával a 49,604%-át lehet elérni.

8. ábra

A döntési fa koncentrációs görbéje



Forrás: saját készítés

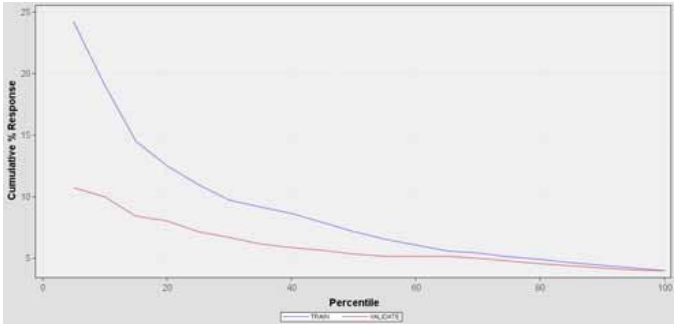
Neurális háló

A neurális háló *Neural Network* elnevezésű eszközzel történő elkészítéséhez a szükséges módosításokat a szűrés és helyettesítés lépéseiben hajtottuk végre, ez esetben az adatok transzformációjára nincs szükség *(SAS 2006)*. A hálóépítés a tanulmány első részében ismertetett módon történt, mely során a felhasznált iterációk maximális számát 10 darabban határoztuk meg, a rejtett rétegben 3 elemet definiáltunk, modellépítési kritériumként pedig az előzőekhez hasonlóan azt a szabályt, miszerint a modell a legnagyobb átlagos profitot, illetve a legkisebb esetleges veszteséget ígérje. Ennek a kritériumnak eleget téve a végső háló az első iteráció után kapott hálóval egyezik meg, mivel az iterációk számának növelésével a tanulóállomány esetén az átlagos nyereség ugyan nő, de ez a megállapítás az érvényesítőállományon már nem állja meg a helyét, a háló további finomítását ezért le kellett állítani.[24]

A teljesítményt bemutató ábrák az előzőekhez hasonló elrendezésben láthatók.

9. ábra

A neurális háló kumulált válaszarány-görbéje



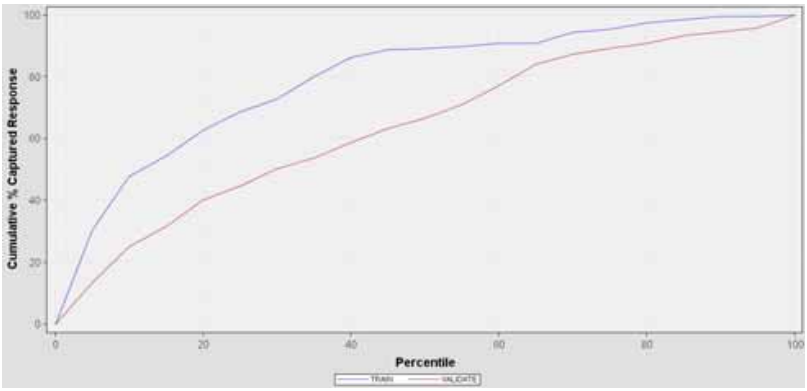
Forrás: saját készítés

[24] A neurális hálóra az érintett ábrákon annak angol megfelelője alapján Neural Network-ként hivatkozunk.

A 9. ábráról leolvasható, hogy a neurális háló becslése alapján a várhatóan pozitívan válaszolók legjobb 10%-ának megkeresésével 9,969%-os, 20%-uk megkeresésével pedig 8,024%-os pozitív válasz érhető el. A kumulált lift-érték azt mutatja, hogy a modell a véletlenszerű megkereséshez képest 10%-os megcélzás esetén 2,492-szeres, 20%-os megcélzásnál pedig 2,006-szeres javulást mutat.
A 10. ábrán látható, hogy a legkedvezőbb várható reakciójú ügyfelek első 10%-ának megkeresése esetén a várhatóan pozitívan reagálók 24,924%-át, az első 20%-uk megkeresése mellett pedig 40,122%-át lehet elérni.

10. ábra

A neurális háló koncentrációs görbéje



Forrás: saját készítés

Együttes modell

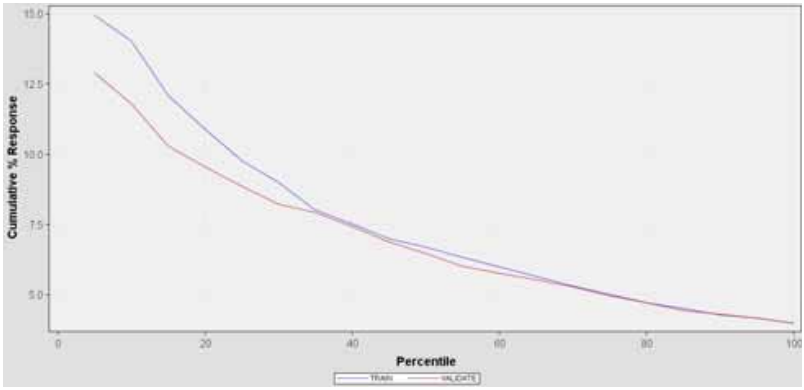
Az *Ensemble* eszköz segítségével az eddigi modellek kombinálása által létrehozhatóvá válik egy azoktól különálló, azok erősségét mégis magán viselő modell felállítása *(SAS 2006)*. Ezt az első regressziós modell, a döntési fa és a neurális háló által becsült pozitív válaszadási valószínűségek átlagolása révén készítettük el, mely valószínűségek az együttes modell becsült válaszadási valószínűségeivé váltak.[25]

A modell teljesítményét az eddigi gyakorlatnak megfelelő ábrák mutatják be.

[25] *Az elkészített modellre az érintett ábrákon Ensemble néven hivatkozunk.*

11. ábra

Az együttes modell kumulált válaszarány-görbéje

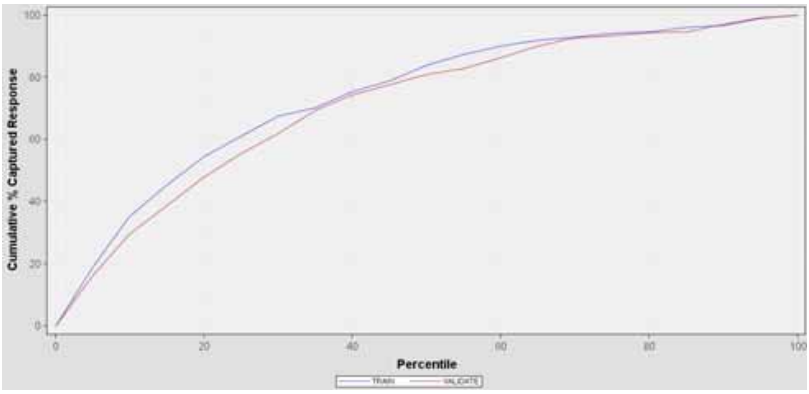


Forrás: saját készítés

A 11. ábráról leolvasható, hogy az együttes modell becslése alapján az ügyfelek legjobb 10%-ának megkeresésével 11,793%-os, 20%-uk megkeresésével 9,544%-os pozitív válasz érhető el. A kumulált lift-érték kiszámításából az látható, hogy a véletlenszerű megkereséshez képest ez 2,948-szeres, illetve 2,386-szeres hatékonyságjavulást jelent.
A 12. ábrán látható, hogy az ügyfelek 10%-ának megcélzása által a várhatóan pozitívan reagálók 29,483%-át, 20%-uk megcélzásával pedig 47,72%-át lehet elérni.

12. ábra

Az együttes modell koncentrációs görbéje



Forrás: saját készítés

Assess mint értékelés

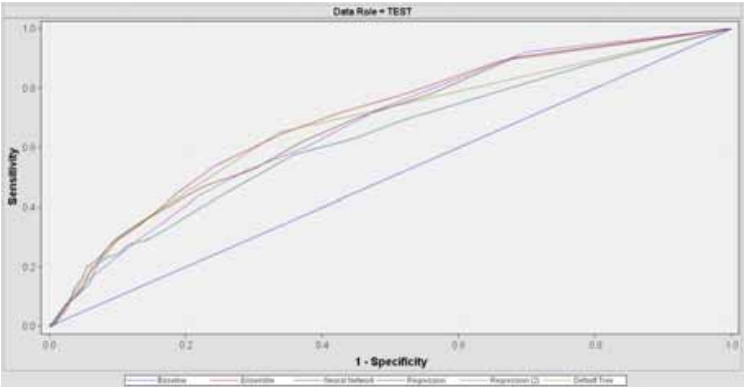
A különböző modellek elkészítése után azok teljesítményének összehasonlítása következhet a *Model Comparison* elnevezésű eszközzel *(SAS 2006)*, melyre alapvetően osztályozási, adatbányászati és statisztikai eszközök alkalmazhatóak. Az esettanulmány szempontjából

az osztályozási értékelő eszközökön belül az úgynevezett ROC-ábrára, az adatbányászati értékelő eszközökön belül a már említett kumulatív lift-értékre, valamint az egy főre jutó átlagos nyereségre helyezzük a hangsúlyt.

A 13. ábrán látható ROC-ábra a modellek becslése alapján hibásan, illetve helyesen előrejelzett pozitív válaszadások között teremt kapcsolatot *(Tan et al. 2005)*.

13. ábra

ROC-ábra



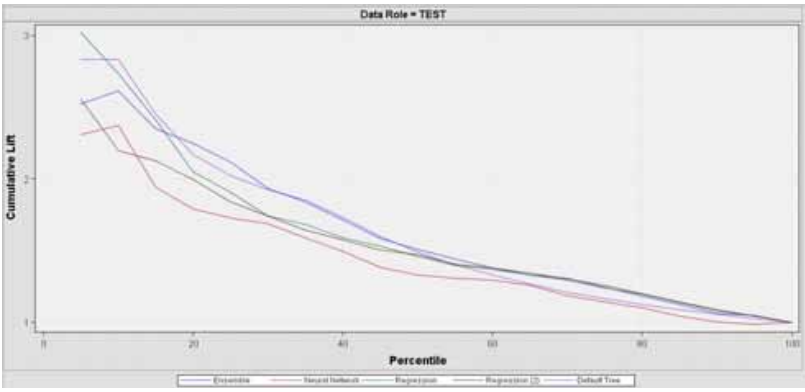
Megjegyzés: Az abszcisszán a modellek becslése alapján hibásan besorolt pozitív válaszok aránya, az ordinátatengelyen pedig a helyesen előrejelzett pozitív válaszadások aránya látható a tesztelőállományra vonatkoztatva a különböző határpontoknak megfelelően.
Forrás: saját készítés

Az ábra kék színnel jelölt 45 fokos átlója a véletlenszerű kiválasztás eredményét képviseli, mely esetén a határpont elmozdítása a helyesen besorolt pozitív válaszok arányának növekedésével megegyező arányú növekményt okoz azok helytelen besorolásában *(Tan et al. 2005)*. Az ábrán együtt szerepel valamennyi elkészített modell. Döntési kritériumként megfogalmazható, hogy minél magasabb egy modell ROC-görbéjének homorúsági foka, tehát minél nagyobb terület található alatta, az annál pontosabb, így jobb becslést képes adni a célváltozó kimenetére vonatkozóan *(Tan et al. 2005)*. Ezt a területet a ROC-indexszel lehet mérni *(SAS 2008)*, ami az együttes modell választását indukálja.

Az értékelés adatbányászati oldala azt mondja ki, hogy a jövőben azt a modellt kell osztályozásra használni, amely a legmegbízhatóbban és legpontosabban képes megbecsülni a célváltozó pozitív kimeneteit, tehát egy tervezett, előre megadott kiválasztási arány mellett a legmagasabb tesztelőállományra számított kumulált lift-értéket biztosítja *(SAS 2006)*. A mutató különböző kiválasztási arányokhoz számított értékeiből képzett görbéje az egyes modellekre a 14. ábrán látható. Mivel a kumulált liftérték azt mutatja meg, hogy a megcélzottak egy adott arányáig egy adott modell mekkora hatékonyságjavulást képes elérni a véletlenszerű kiválasztáshoz képest *(Coppock 2002)*, ezért azt a modellt kell választani, melynek tesztállományra vonatkoztatott görbéje a kívánt kiválasztási aránynál a legmagasabban helyezkedik el.

14. ábra

A modellek kumulált liftérték görbéi



Forrás: saját készítés

A 14. ábra alapján elkészített, a megcélozni kívánt ügyfelek egy adott arányánál alkalmazandó leghatékonyabb osztályozási modellt a 2. táblázat mutatja be.

2. táblázat

A hatékonyságjavulás szempontjából alkalmazandó modell

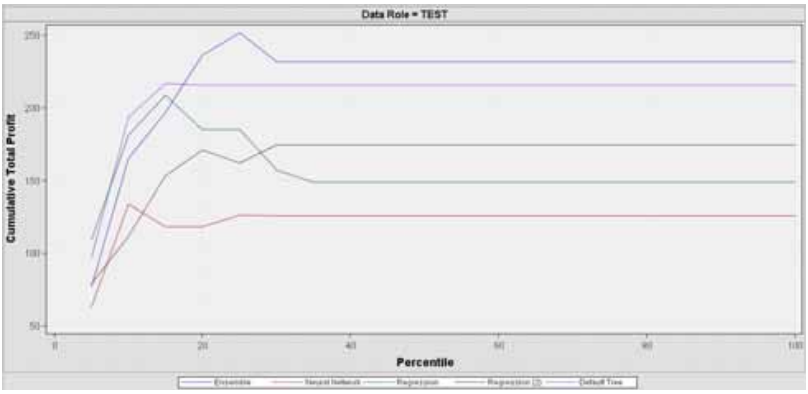
Megcélzott ügyfelek százalékos aránya	Alkalmazandó modell
0 – 8,3%	Első regressziós modell
8,3 – 18%	Döntési fa
18 – 31,6%	Együttes modell
31,6 – 47,3%	Döntési fa
47,3 – 60%	Együttes modell
60 – 100%	Második regressziós modell

Forrás: saját készítés

Az osztályozási minőséget és az adatbányászati szempontokat érvényesítő értékelési szempontok mellett azonban az üzleti érdekeket is figyelembe kell venni az alkalmazandó modell kiválasztásánál. Erre a legmegfelelőbb mutató a kumulált teljes profit *(SAS 2006)*, melynek az egyes modellekre vonatkozó görbéit a 15. ábra tartalmazza.

15. ábra

A modellek kumulált teljes profit görbéi



Forrás: saját készítés

Az ügyfelek különböző megkeresési arányai mellett az egyes modellek más-más profitot ígérnek, így ennek függvényében történhet az alkalmazandó modell kiválasztása. A 15. ábra alapján elkészített, a megcélozni kívánt ügyfelek egy adott arányánál alkalmazandó profitmaximalizáló modellt a 3. táblázat mutatja be.

3. táblázat

A profitmaximalizálás szempontjából alkalmazandó modell

Megcélzott ügyfelek százalékos aránya	Alkalmazandó modell
0 – 7,6%	Első regressziós modell
7,6 – 17,7%	Döntési fa
17,7 – 100%	Együttes modell

Forrás: saját készítés

Amennyiben a vállalat stratégiája egy szélesebb ügyfélkör megcélzását is megköveteli, akkor egy modell jóságának az egész tartományon fenn kell állnia.[26] Ebben az esetben az egyik lehetséges modellválasztási kritérium az egy főre jutó átlagos profit *(SAS 2006)*. Ezeket az átlagos profitokat a 4. táblázat tartalmazza, mely alapján megállapítható, hogy ez esetben az együttes modellt kell alkalmazni a nyereség maximalizálása érdekében.

[26] Erre példa egy többkategóriás kampány, mely során a legesélyesebb ügyfeleket költségesebb, de hatékonyabb csatornákon – például papíralapú levél útján – keresik meg, míg a közepesen ígéretes ügyfeleket költségkímélőbb eszközök – például elektronikus levél – által.

4. táblázat

A modellek által ígért átlagos nyereségek

Modell neve	Modell által ígért egy főre jutó profit
Neurális háló	0,05592 euró/fő
Első regressziós modell	0,07388 euró/fő
Második regressziós modell	0,07492 euró/fő
Döntési fa	0,09577 euró/fő
Együttes modell	0,10301 euró/fő

Forrás: saját készítés

Összegzésként elmondható, hogy mind az osztályozás minősége, mind az üzleti szempontok érvényesülése szempontjából az együttes modell választandó.[27]

A modellek értékelése után következhet az adatbányászati projekt záró lépése, az úgynevezett pontozás. Ez egy olyan folyamatot takar, mely eredményeként az eljárás új, célváltozóval nem rendelkező adatállományokra is alkalmazhatóvá és más alkalmazási környezetben is használhatóvá válik. A pontozás során az egyes inputváltozók és azok különféle transzformációi jelentőségük alapján pontozásra kerülnek, mely segítségével minden egyes ügyfélre létrehozhatóvá válik annak becsült válaszadási valószínűsége. Ezen valószínűségek meghatározásával könnyen definiálhatóvá válnak azok az ügyfelek, akiket egy direktmarketing-kampány során ajánlott megkeresni, mivel ők azok, akik bizonyíthatóan a legvalószínűbben fognak válaszolni *(SAS 2008)*.

A pontozás eredményét egy újabb adatállomány hiányában az eredeti 7500 megfigyelésből álló adatállományra alkalmaztuk oly módon, hogy abból eltávolítottuk a célváltozó oszlopát. Az eljárás eredményeként meghatározásra kerültek az egyes ügyfelek pozitív válaszadásának becsült valószínűségei, melyek alapstatisztikáit az 5. táblázat tartalmazza.

5. táblázat

Az ügyfelek becsült válaszadási valószínűségei

	Becsült valószínűség
Átlag	0,05
Szórás	0,05
Minimum	0,01
Első kvartilis	0,01
Medián	0,03
Harmadik kvartilis	0,06
Maximum	0,54

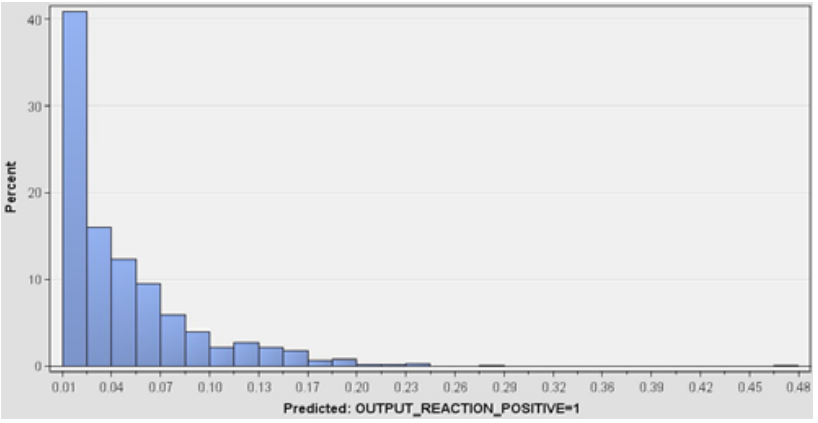
Forrás: saját készítés

[27] Előző szempontok mellett azonban a modellek realitásának megőrzése érdekében ügyelni kell az azok pontosságát jellemző statisztikákra is, például a téves osztályozási rátára.

Az egyes ügyfelek pozitív válaszadásának becsült valószínűségeiből készített hisztogramot a 16. ábra tartalmazza.

16. ábra

A pozitív reakciók valószínűségének hisztogramja



Forrás: saját készítés

Az ábráról leolvasható, hogy az egyes becsült valószínűségekhez az ügyfelek mekkora hányada tartozik. Az ábra jobb oldalán szerepelnek a legjobb ügyfelek, így ha például a legígéretesebb 25%-ot kívánjuk megcélozni, akkor azokat az ügyfeleket kell kiválasztani, melyek pozitív válaszadásának becsült valószínűsége legalább 0,06. Ezen információ birtokában már könnyen elkészíthető egy olyan jelentés, mely az eredeti adatbázist a gyakorlatban használható tudássá alakítva tartalmazza a megkeresésre kijelölt címzettek listáját.

Következtetések

A tanulmány első részében áttekintett módszertani ismereteknek a második részben bemutatott gyakorlati alkalmazása egyértelműen alátámasztja azt a tényt, mely szerint az adatbányászat üzleti keretek között történő felhasználása több szempontból is előnyös lehet a vállalatok számára. Ezt igazolja a felépített modellek által feltárt tudásban rejlő lehetőségek sokasága is. Elég itt csupán arra gondolni, hogy az elemzés során megszerzett információk a hatékonyságnövekedés elérése mellett időmegtakarítást, illetve az előzőekből következően a bevételek növelését és a költségek racionalizálását teszik lehetővé. Ezek a pozitív következmények pedig nem korlátozódnak a direktmarketing területére, hanem számos egyéb alkalmazási területen is realizálhatóak. Ráadásul jellemzően oly mértékben valósíthatóak meg, hogy hatásuk releváns módon képes megnyilvánulni a vállalat eredményességében is.

Az eredmények általánosításával megállapítható, hogy az adatbányászat reális körülmények között történő felhasználása a profitorientált piaci szereplők hatékonyságának és eredményességének növelését teszi lehetővé. Így annak figyelembevételével, hogy az alkalmazásához szükséges szoftverberuházás és szakértői gárda – kiszervezés esetén pedig az igénybevett szolgáltatás – óriási összegeket emészthet fel, a bevezetés egy kellően

nagyméretű, megfelelő pénzügyi háttérrel rendelkező és innovatív vállalat számára többszörösen megtérülő beruházást jelenthet. A megtérülés itt nem csupán pénzügyi értelemben értendő, hiszen a feladatok és az azokkal szemben támasztott követelmények átláthatóbbá, elvégzésük tudatosabbá, eredményük látványosabbá válik. Az említett hatások hosszú távon a munkavállalók motivációjában és lojalitásában is tetten érhetők, azonban ezek már túlmutatnak a tanulmány eredeti témáján, így elemzésüket az érintett szakterület kutatóira hagyjuk.

Hivatkozások

Ary Bálint Dávid – Dr. Imre Sándor [2006]: *Számlázás újgenerációs telekommunikációs hálózatokban*. Híradástechnika, LXI. évfolyam, 10. szám, 40–45. o.

Berry, M. J. A. – Linoff, G. [1997]: *Data Mining Techniques: For Marketing, Sales, and Customer Support*. John Wiley and Sons, Inc., New York.

Coppock, D. S. [2002]: *Data Modeling and Mining: Why Lift?* In: http://www.information-management.com/news/5329-1.html (Letöltve: 2009. február 21.)

Fisher, R. A. [1915]: *Frequency Distribution of the Values of the Correlation Coefficient in Samples of an Indefinitely Large Population*. Biometrika, vol. 10. no. 4. 507–521. o.

Hunyadi László – Vita László [2004]: *Statisztika közgazdászoknak* (Harmadik átdolgozott kiadás). KSH, Budapest.

Lucas, A. [2004]: *The Gini Coefficient*. In: http://www.rhinorisk.com/Publications/Gini%20Coefficients.pdf (Letöltve: 2009. március 09.)

Márkus Béla [1994]: *Térinformatika* (Főiskolai jegyzet). Erdészeti és Faipari Egyetem, Földmérési és Földrendezői Főiskolai Kar, Székesfehérvár. 23. o.

MNB [2008]: *Lekérdezhető árfolyamok.* In: http://www.mnb.hu/engine.aspx?page=arfolyamtablazat&query =2008.11.26.,2008.11.26.,1,EUR (Letöltve: 2008. december 12.)

SAS [2006]: *Introduction to SAS® Enterprise Miner™ Course Notes*. SAS Institute INC., Cary, NC.

SAS [2008]: *Getting Started with SAS® Enterprise Miner™ 5.3*. SAS Institute Inc., Cary, NC. In: http://support.sas.com/documentation/onlinedoc/miner/getstarted53.pdf (Letöltve: 2009. február 12.)

Tan, P. N. – Steinbach, M. – Kumar, V. [2005]: *Introduction to Data Mining*. Addison-Wesley, Richmond, TX.